Exercise 3

Juuso Korhonen - 652377
ELEC-E8125 - Reinforcement Learning

October 3, 2022

1 Task 1.1

Implement Q-learning as presented in [1] Section 6.5 for the Cartpole environment in file train.py . We need to compare two exploration methods:

(a) using a constant value of $= 0.1$;

(b) using GLIE (i.e. greedy in limit with infinite exploration) which reduces the value of over time. Its formula can be found from the lecture.

For GLIE, aim at reaching $= 0.1$ after 20000 episodes and round the value of constant b to the nearest integer.

1.1 a

For constant epsilon case, training performance plots for episode reward is in figure 1, and for smoothed episodic reward in figure 2.

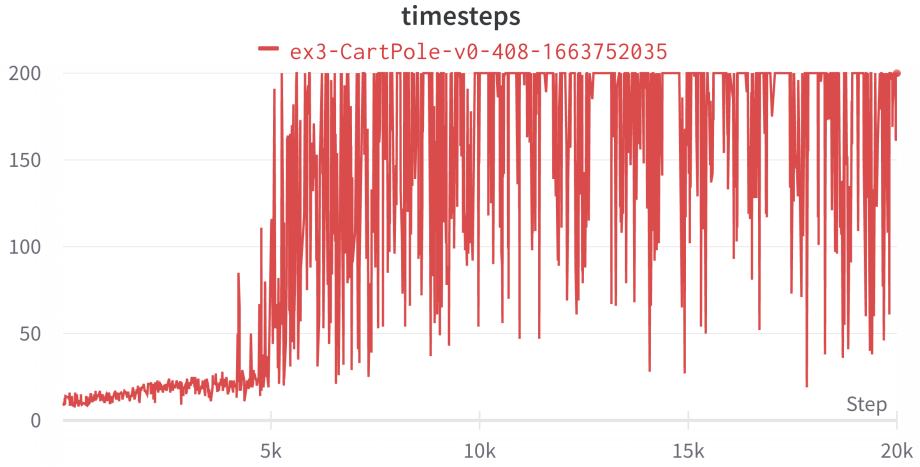

Figure 1: Train episode reward with constant epsilon.

For GLIE case, training performance plots for episode reward is in figure 3, and for smoothed episodic reward in figure 4.

Figure 2: Smoothed train episode reward with constant epsilon.

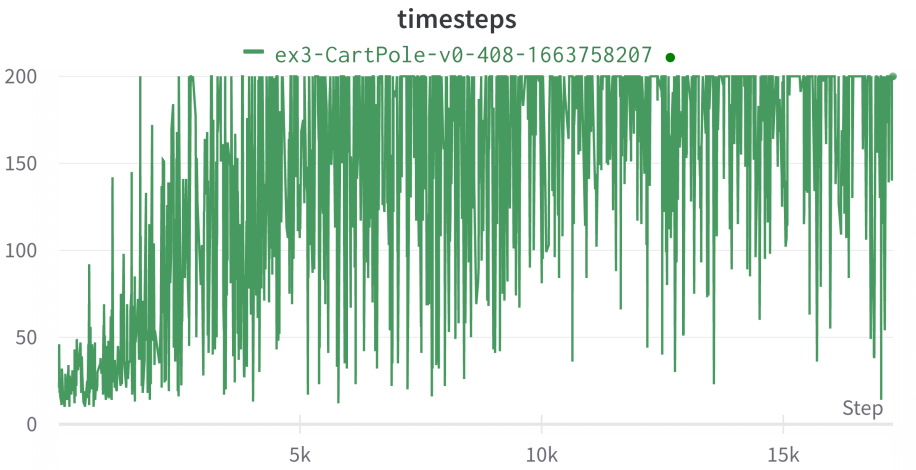

Figure 3: Train episode reward with GLIE.

Figure 4: Smoothed train episode reward with GLIE.

2 Task 1.2

Heatmap of the value function estimated from the Q-function with GLIE is plotted in figure 5.

2.1 Question 1

What do you think the heatmap would have looked like:

(a) before training?

It would have been zeros everywhere.

(b) after single episode?

I expect there to be mostly zeros, but some low values around the starting position due to high initial exploration.

(c) halfway through the training?

I expect the explored states to be close to the same as after training, but the values to be lower/more evenly distributed since they are not yet well defined (they are defined better in the greedy phase of GLIE).

2.2 Task 1.3

Set ϵ to zero, effectively making the policy greedy w.r.t. current Q-value estimates. Run the training again while:

(a) keeping the initial estimates of the Q function at 0

The training performance plots for (a) case are in figures 6 and 7.

(b) setting the initial estimates of the Q function to 50 for all states and actions.

The training performance plots for (b) case are in figures 8 and 9.

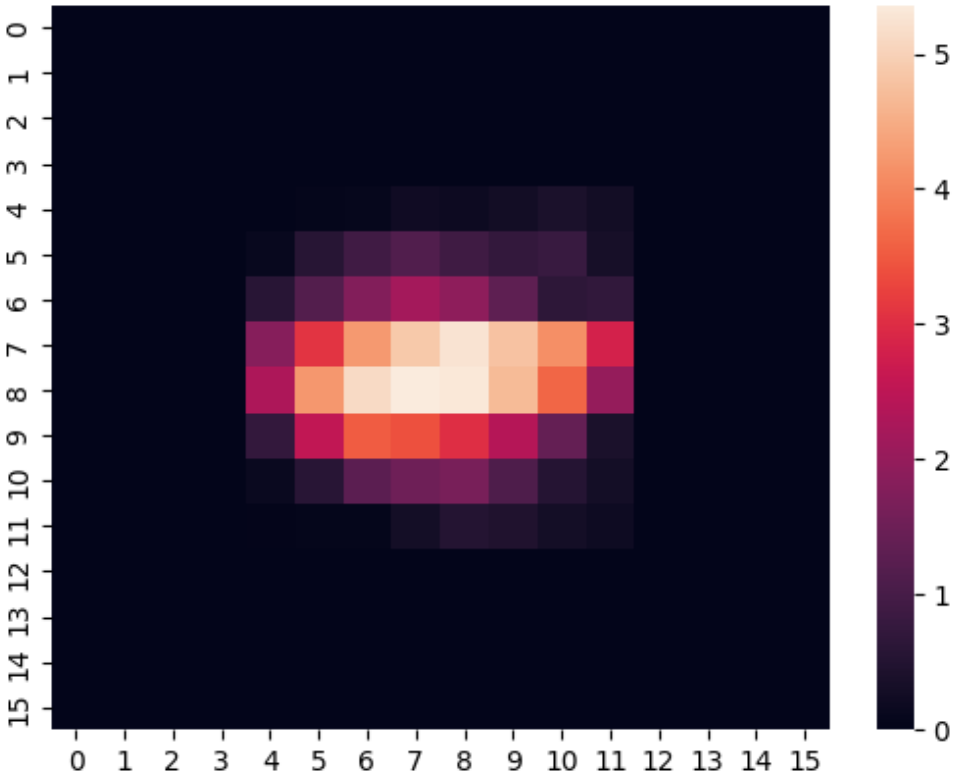

Figure 5: Heatmap of the value function estimated from the Q-function with GLIE.

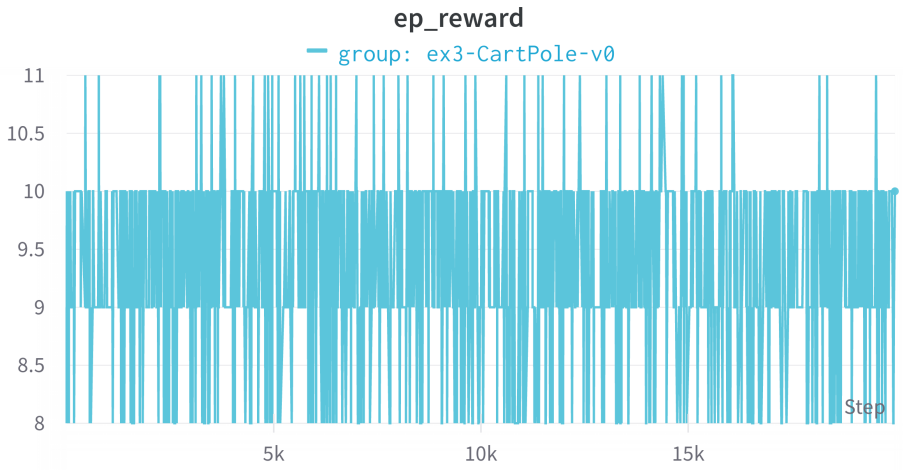

Figure 6: Train episode reward with ϵ and initial Q-values set to zero.

Figure 7: Smoothed train episode reward with $\epsilon = 0$ and initial Q-values set to zero.

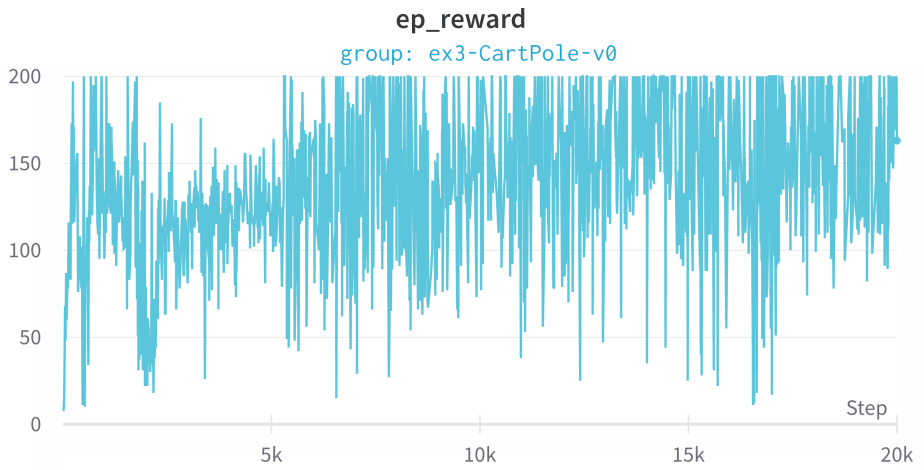

Figure 8: Train episode reward with $\epsilon = 0$ and initial Q-values set to 50.

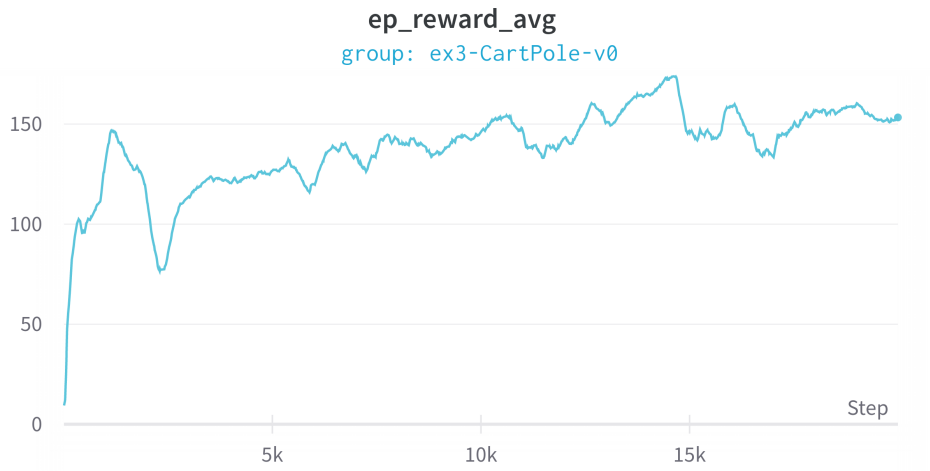

Figure 9: Smoothed train episode reward with $\epsilon = 0$ and initial Q-values set to 50.

3 Question 2

3.1 Question 2.1

Model with the initial q-values at 50 performed better.

3.2 Question 2.2

Setting high initial Q-values leads to more exploration as the initial Q-value-updates lead to reduction if the reward is not high enough to cover the discount on the initial value ($R < (1 - \gamma)Q(0)$).

4 Task 2

Training performance plots for Lunar Lander are plotted in figures 10 and 11.

4.1 Question 3

Q: Does the lander learn to land between the flag poles? Why/why not?

The lander learns to sometimes land between the flag poles, but most of the time it does not manage to do that. This is likely due to increase in state-action space, and 20,000 iterations of Q-learning not being enough to cover it well.

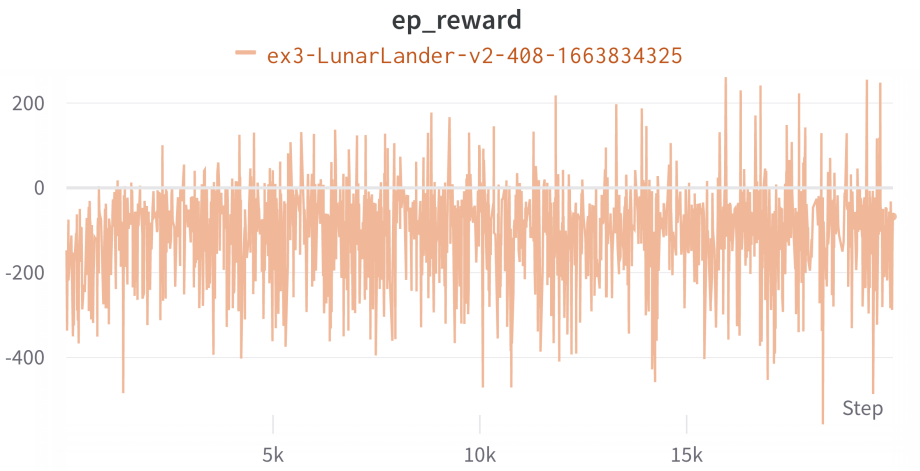

Figure 10: Train episode reward for Lunar Lander environment.

Figure 11: Smoothed train episode reward for Lunar Lander environment.